Tight Bounds on Vertex Connectivity Under Vertex Sampling

Keren Censor-Hillel, Mohsen Ghaffari, George Giakkoupis, Bernhard Haeupler, Fabian Kuhn

Tight Bounds on Vertex Connectivity Under Vertex Sampling

Keren Censor-Hillel* Mohsen Ghaffari† George Giakkoupis‡ Bernhard Haeupler§
Fabian Kuhn¶

Abstract

A fundamental result by Karger [10] states that for any λ-edge-connected graph with n nodes, independently sampling each edge with probability $p = \Omega(\log n/\lambda)$ results in a graph that has edge connectivity $\Omega(\lambda p)$, with high probability. This paper proves the analogous result for vertex connectivity, when sampling vertices. We show that for any k-vertex-connected graph G with n nodes, if each node is independently sampled with probability $p = \Omega(\sqrt{\log n/k})$, then the subgraph induced by the sampled nodes has vertex connectivity $\Omega(kp^2)$, with high probability. This bound improves upon the recent results of Censor-Hillel et al. [6], and is existentially optimal.

1 Introduction

Consider a random process where given a base graph G, each edge or node of G is sampled with some probability p. Given such a random graph process, it is interesting to see how various global connectivity properties of the graph induced by the sampled edges or nodes change as a function of the sampling probability p. If G is the complete n-node graph, sampling each edge independently with probability p results in the classic Erdős-Rényi random graph $G_{n,p}$, for which exact thresholds for the formation of a giant component, global connectivity, and many other properties have been studied (e.g., in [3]). Thresholds for the formation of a giant component are further studied more generally in percolation theory [4]—mostly for graphs G defined by some regular or random lattice. In the context of percolation theory, edge sampling is called bond percolation whereas vertex sampling is referred to as site percolation.

In the present work, we are interested in how the vertex connectivity changes under uniform random vertex sampling of a general n-vertex graph G. For edge connectivity and edge sampling, the analogous question has been resolved two decades ago. Karger's seminal result [10] showed that for any λ-edge-connected graph with n vertices, sampling edges independently at random with probability $p = \Omega(\log n/\lambda)$ results in an $\Omega(\lambda p)$-edge-connected subgraph, with high probability[1]. This was a strong extension of the earlier result by Lomonosov and Polesskii [13], which stated that sampling each edge with probability $\Theta(\log n/\lambda)$ leads to a connected subgraph, w.h.p. These sampling results and their extensions were cornerstone tools for addressing various important problems such as various min-cut problems [9, 10], constructing cut-preserving graph sparsifiers [2, 15], max-flow problems [9, 12], and network reliability estimations[11].

As in the case of edge connectivity, studying the vertex connectivity of the subgraph obtained by independently sampling vertices of a k-vertex-connected graph is of fundamental interest. However, progress in this problem has been much scarcer. Up to a year ago, it was not even known whether a $\Theta(n)$-vertex-connected graph stays (simply) connected when nodes are sampled with probability $p = 1/2$. Recently, Censor-Hillel et al. [6] showed that a sampling probability of $p = \Omega(\log n/\sqrt{k})$ is a sufficient condition for connectivity (w.h.p.), and moreover, it was proven that the remaining vertex connectivity of the sampled subgraph is $\Omega(kp^2/\log^3 n)$, w.h.p. However, it remained open whether these two bounds are optimal.

In this paper, we answer this question by providing tight bounds.

THEOREM 1. *Let $G = (V, E)$ be a k-vertex-connected n-node graph, and let S be a randomly selected subset of V where each node $v \in V$ is included in S independently with probability $p \geq \alpha\sqrt{\log n/k}$, for a sufficiently large constant α. Then the subgraph $G[S]$ of G induced by S is connected with high probability. Moreover, $G[S]$ has vertex connectivity $\Omega(kp^2)$, with probability $1 - e^{-\Omega(kp^2)}$.*

Theorem 1 improves over [6] in two ways: its first part improves over [6, Theorem 1.7] which only proves connectivity for a sampling probability $p = \Omega(\log n/\sqrt{k})$; its second part improves over [6, Theorem 1.4], which proves a remaining vertex connectivity of $\Omega(kp^2/\log^3 n)$.

In the rest of this section, we first give a brief explanation of why the standard techniques used for the edge con-

*Shalon fellow at Technion, Israel, ckeren@cs.technion.ac.il
†MIT, USA, ghaffari@mit.edu
‡INRIA Rennes, France, george.giakkoupis@inria.fr
§Carnegie Mellon U., USA, haeupler@cs.cmu.edu
¶U. of Freiburg, Germany, kuhn@cs.uni-freiburg.de

[1] We use the phrase '*with high probability*' (*w.h.p.*) to indicate that some event has a probability of at least $1 - n^{-\Theta(1)}$.

nectivity case do not work for vertex connectivity, and we present a brief explanation of our approach and how it compares with that of [6]. Then we discuss a simple graph construction that shows the optimality of the bounds in Theorem 1, and finally, we state some implications of Theorem 1.

1.1 Overview of the Proof

The Challenge. To understand the challenge, we briefly explain why tools with a similar flavor to the ones used for edge connectivity do not take us far in the vertex connectivity case. The key to most results about edge sampling is the "cut counting" argument introduced in [8], where it is shown that in a graph of edge connectivity λ, the number of cuts of size at most $\alpha\lambda$ is at most $n^{O(2\alpha)}$. Combined with a standard Chernoff argument and a union bound over all cuts, this shows that when independently sampling each edge with probability $p = \Omega(\log n/\lambda)$, it holds w.h.p. for the subgraph induced by the sampled edges, that the size of each cut does not deviate from its expectation by more than a constant factor [10]. Hence, in particular, the edge connectivity of the sampled subgraph is $\Omega(\lambda p)$, w.h.p. Unfortunately, the same approach cannot work for vertex sampling and vertex connectivity, because in graphs with vertex connectivity k, even the number of minimum vertex cuts can be as large as $\Theta(2^k(n/k)^2)$ [7].

The Old Approach. In [6], the bound on the sampling threshold for (simple) connectivity is obtained by essentially considering the vertex sampling as a gradual process that happens in phases, and by analyzing the growth of the connected components throughout this process. More precisely, it is shown that when starting from a dominating set,[2] if each node is sampled with probability $1/\sqrt{k}$, then in expectation, the number of connected components drops by a constant factor. Hence, after $O(\log n)$ phases where in each phase nodes are sampled independently with probability $1/\sqrt{k}$, and thus after an overall sampling probability of $O(\log n/\sqrt{k})$, the subgraph induced by the sampled nodes is connected, w.h.p.

This gradual process is not sufficient on its own for proving values of vertex connectivity higher than one. To prove higher remaining vertex connectivity while trying to avoid explicitly working on all cuts, [6] developed the notion of connected dominating set (CDS) packings. This notion serves as a certificate for large vertex connectivity (among other applications). Particularly, it is shown that after sampling with probability p, it is possible to construct a fractional CDS packing of size $\Omega(kp^2/\log^3 n)$. Since the size of any (fractional) CDS packing of a graph is upper bounded by its vertex connectivity, this directly implies that the vertex connectivity of the remaining graph is also at least $\Omega(kp^2/\log^3 n)$. While two of the logarithmic factors in this approach seem to be artifacts of the details in the method, the third one appears to be an inherent limitation of the method. This is because [6] shows that there are graphs with vertex connectivity k that have maximum (fractional) CDS packing size of $O(k/\log n)$. Thus, the approach of using a CDS packing as a witness for the vertex connectivity of the sampled subgraph inherently cannot prove a bound better than $\Omega(kp^2/\log n)$.

The New Approach. In this paper, we provide a completely new approach for analyzing the remaining vertex connectivity after vertex sampling. Instead of sampling with probability $1/\sqrt{k}$ in each phase, we sample with a lower probability of $1/k$. In addition, rather than directly showing that the sampled nodes induce a connected subgraph, we introduce the notion of λ-*semi-connectivity* which allows us to analyze the progress in a more refined way. We call a vertex set $S \subseteq V$ λ-semi-connected if for every partition of the connected components of the induced subgraph $G[S]$ into two parts, there are λ nodes in $V \setminus S$ which are adjacent to components on both sides of the partition. Roughly, in each phase, each component of $G[S]$ in expectation gains at least one new such connector node to another component, and we use this to show that within $t+\log n$ phases, the semi-connectivity of the sampled set of nodes grows by $\Omega(t)$. It follows that when sampling with probability p, we get semi-connectivity at least $\Omega(kp)$. This is the main technical contribution of our paper, and it is shown by carefully analyzing how the semi-connectivity of the sampled set grows by adding new random vertices. Having now a set with semi-connectivity $\Omega(kp)$, we can use techniques similar to the ones used in [1, 6] to show that another round of sampling with probability p yields a connected graph with probability $1 - e^{-\Theta(kp^2)}$ (as long as $kp^2 = \Omega(\log n)$).

To show that the remaining vertex connectivity after sampling is large, we exploit the fact that the probability for the graph to not be connected is exponentially small in kp^2. Suppose we want to show that the vertex connectivity after sampling is at least $k' + 1$ for some $k' \ll k$. Given any $X \subset V$ of k' nodes, the graph $G[V \setminus X]$ induced by $V \setminus X$ has vertex connectivity at least $k - k' \approx k$, and thus sampling each vertex of $G[V \setminus X]$ with probability p yields a connected subgraph with probability $1 - e^{-\Theta(kp^2)}$. A union bound over all $\binom{n}{k'} = e^{O(k' \log n)}$ k'-node subsets of V implies that the subgraph after sampling G is at least $(k' + 1)$-vertex-connected, for some $k' = \Omega(kp^2/\log n)$. A conceptually similar, but more careful argument shows that the vertex connectivity of the sampled subgraph is in fact $\Omega(kp^2)$.

[2] A graph with vertex connectivity k has minimum degree at least k, and thus a dominating set is already obtained w.h.p. when sampling with probability $\Omega(\log n/k)$.

1.2 Optimality of our Results. The bounds in Theorem 1 are existentially tight up to constant factors, as demonstrated in the following simple example.[3]

OBSERVATION 2. *Let G be a $2n$-node graph consisting of two disjoint n-node cliques connected via a matching of $k \leq n$ edges. The vertex connectivity of G is k, and when each node is sampled with probability $p \geq 2\ln n/n$, the expected vertex connectivity of the subgraph induced by the sampled nodes is at most $kp^2 + o(kp^2)$. If the sampling probability is $p = o(\sqrt{\log n/k})$, then the subgraph is disconnected[4] with probability at least $1/n^{o(1)}$.*

Even if one desires the sampled subgraph to be connected with merely a *constant* probability, our sampling threshold $p = \Omega(\sqrt{\log n/k})$ is essentially tight as shown by the next simple example.

OBSERVATION 3. *Let G be an n-node graph consisting of n/k k-cliques ordered 1 to n/k, where each two consecutive cliques are connected via a k-edge matching. We assume that n is a multiple of k, and $k < n$. Graph G has vertex connectivity k, and when sampling nodes with probability $p = o(\sqrt{\log(n/k)/k}) \cap \omega(1/n)$, the subgraph induced by the sampled nodes is disconnected with probability $1 - o(1)$.*

1.3 Implications. The fact that Theorem 1 proves an $\Omega(k)$ remaining vertex connectivity when $p = 1/2$, combined with the approach in [6, Section 5], imply the following corollary.

COROLLARY 4. *Any k-vertex-connected n-node graph can be decomposed into $\Omega(k/\log^2 n)$ vertex-disjoint connected dominating sets (CDS).*

This improves over the $\Omega(k/\log^5 n)$ bound of [6, Theorem 1.2]. As explained in [6, 5], decomposing to vertex-disjoint connected dominating sets can be viewed as a decomposition of vertex-connectivity. This makes Corollary 4 the best known counterpart of the famous results of Tutte [16] and Nash-Williams [14] from 1961 for decomposing edge-connectivity; namely that each λ-edge-connected graph contains $\lceil \frac{\lambda-1}{2} \rceil$ edge-disjoint spanning trees. The $\Omega(k/\log^2 n)$ bound of Corollary 4 is within an $O(\log n)$ factor of optimal because, as shown in [6], there exist k-connected graphs that cannot be decomposed into more than $\Theta(k/\log n)$ vertex disjoint connected dominating sets. Furthermore, the decomposition stated in Corollary 4 can be computed very efficiently, namely in $\tilde{O}(m)$ time where m is the number of edges in the graph, by combining random sampling with the approach of [5].

[3] The proofs of Observation 2 and 3 are given in Section 3.

[4] For the purposes of this statement, we consider the empty graph disconnected.

In addition, following the connection stated in [5, Section 1.4.1], Corollary 4 implies the best known approximation of the 1989 conjecture of Zehavi and Itai [17]. This conjecture states that each k-vertex-connected graph contains k *vertex-independent trees*, that is, k spanning trees rooted in a node $r \in V$ such that for each vertex $v \in V$, the paths between r and v in different trees are internally vertex-disjoint. We get the following approximation.

COROLLARY 5. *Any k-vertex-connected n-node graph contains $\Omega(k/\log^2 n)$ vertex-independent trees.*

2 Proof of Theorem 1

The main part of our analysis focuses on proving the following result, which is a refinement of the first part of Theorem 1. It provides a lower bound for the sampling probability p that maintains connectivity, that is a function also of the probability $1 - \delta$ that $G[S]$ is connected.

THEOREM 6. *Let $G = (V, E)$ be a k-vertex-connected n-node graph. For an arbitrary $0 < \delta < 1$ (that can be a function of n and k), let S be a randomly selected subset of V such that each $v \in V$ is included in S independently with probability $p \geq \beta\sqrt{\log(n/\delta)/k}$, for a sufficiently large constant β. Then the subgraph $G[S]$ induced by S is connected with probability at least $1 - \delta$.*

We show that proving $G[S]$ is connected with very high probability is directly sufficient in showing that $G[S]$ has a large remaining vertex-connectivity. In particular, our next proof gives a generic black-box reduction which proves the second part of Theorem 1 using only Theorem 6. While the proof is quite simple, we believe this reduction to be an important contribution of this paper because it provides a new way to prove higher remaining vertex connectivity. In comparison, as mentioned before, [6] also showed a bound of $\Omega(kp^2/\log^3 n)$ on the remaining vertex connectivity, by constructing a fractional packing of connected dominating sets in the resulting graph, which acts as a certificate for the remaining vertex connectivity. However, as discussed in Section 1.1, this approach cannot prove a remaining connectivity beyond $O(kp^2/\log n)$. We remark that the concentration bound one can get out of the approach of [6] cannot prove a remaining connectivity of more than $\Omega(\sqrt{kp^2/\log^2 n})$, which is extremely weaker.[5]

[5] Note that an $\Omega(\sqrt{kp^2/\log^2 n})$ bound on the remaining vertex connectivity can also be proven in a simpler and more direct way just by randomly throwing each sampled node into one of $\Omega(\sqrt{kp^2/\log^2 n})$ classes, and noticing that each class is a CDS, with high probability, thus implying an $\Omega(\sqrt{kp^2/\log^2 n})$ remaining vertex connectivity.

2.1 Proof of Theorem 1 assuming Theorem 6. Let S be the set of sampled vertices. We want to prove that the subgraph $G[S]$ induced by S is $k'+1$ vertex-connected for some $k' = \Omega(kp^2)$, with probability $1 - 2^{-\Omega(k')}$. We do this by showing that with probability $1 - 2^{-\Omega(k')}$, $G[S]$ is such that it stays connected even if an adversary removes up to k' of its nodes. Actually, we give even more power to this adversary. Simultaneous with the process of random sampling of S, we run an auxiliary random experiment where we randomly color each node in S using a color that is uniformly picked out of $100k'$ colors. Then, we allow the adversary to choose k' colors and remove any subset of nodes of these colors. Let $\mathcal{E}_1$ be the event that there are k' nodes that their removal disconnects $G[S]$, and let $\mathcal{E}_2$ be the event that there are k' colors that removing a subset of nodes of these colors disconnects $G[S]$. Note that $\mathcal{E}_1$ obviously implies $\mathcal{E}_2$. Having this in mind, to show that $\Pr[\mathcal{E}_1] \leq 2^{-\Omega(k')}$, we show that $\Pr[\mathcal{E}_2] \leq 2^{-\Omega(k')}$.

For a set Q of k' colors among the $100k'$ colors, we say that Q is *bad* for set S if removing *some* vertices with colors in the set Q from $G[S]$ disconnects it. To prove that $\Pr[\mathcal{E}_2] \leq 2^{-\Omega(k')}$, we argue that the probability that there is a set Q of k' colors that is bad for S is at most $2^{-\Omega(k')}$. We first fix such a set Q and show that the probability that Q is bad for S is at most $2^{-20k'}$. Then, we use a union bound over all choices of Q to conclude the proof.

Fix an arbitrary set Q of k' colors among the $100k'$ colors. Slightly abusing the notations, let us use $S \setminus Q$ to indicate the set of sampled nodes which have colors other than those in Q. We show that with probability $1 - 2^{-20k'}$, the set $S \setminus Q$ is a connected dominating set of the graph G. Thus, adding any extra node to $S \setminus Q$ also leads to a connected induced subgraph. Hence, this implies that the probability that Q is bad for S is at most $2^{-20k'}$.

First, we show that with probability $1 - 2^{-\Omega(kp)}$, the set $S \setminus Q$ is a dominating set of G. Note that for the fixed set Q, for each node to be in set $S \setminus Q$, it has to be sampled and then not colored by one of the k' colors of Q. Thus, each node is in $S \setminus Q$ with probability $99p/100$ and this decision is independent among different nodes. Since each node in a k-vertex-connected graph has at least k neighbors, we get that the probability that $S \setminus Q$ is not dominating is at most $n(1 - 99p/100)^k = 2^{-\Omega(kp)}$.

Second, we use Theorem 6 while setting $\delta = n/2^{k(\frac{99}{100}p)^2/\beta^2}$, which corresponds to sampling probability of $99p/100$, that is, exactly the probability that each node is in $S \setminus Q$. From Theorem 6, we get that the probability that $G[S \setminus Q]$ is connected is at least $1 - \delta$. For a sufficiently large constant α in Theorem 1, we get $\delta = n/2^{k(\frac{99}{100}p)^2/\beta^2} < 2^{-kp^2/2\beta^2}$. Setting $k' = kp^2/50\beta^2$, this means that $G[S \setminus Q]$ is connected with probability at least $1 - 2^{-25k'}$.

Taking a union bound over the failure of domination and the failure of connectivity argued in the above two paragraphs respectively, we get that the probability that $S \setminus Q$ is not a connected dominating set of G is at most $2^{-25k'} + 2^{-\Omega(kp)} \leq 2^{-20k'}$. Thus, the probability that Q is bad for S is at most $2^{-20k'}$. Now using a union bound over the $\binom{100k'}{k'} \leq (e \cdot 100)^{k'} < 2^{10k'}$ choices for Q, we have that the probability there exists a set of k' colors that is bad for S is less than $2^{-10k'}$. Thus, with probability at least $1 - 2^{-10k'}$, there is no k'-subset of colors that is bad for S. Hence, with probability at least $1 - 2^{-10k'}$, S is such that removing any subset of nodes of k' colors, and thus also any k' nodes, leaves $G[S]$ connected. This completes the proof.

2.2 Intuition for the Proof of Theorem 6. Next, we provide some brief intuition for the approach we use to prove Theorem 6. We explain how we show that a sampling probability of $p = \Theta(\sqrt{\log n/k})$ leads to a connected sampled subgraph w.h.p. This corresponds to the first part of Theorem 1, or equivalently, to Theorem 6 with $\delta = n^{-\Theta(1)}$.

We use a natural interpretation of sampling that was introduced for this problem in [6], in which one looks at the sampling process as slowly adding nodes over time. In particular, instead of sampling nodes with probability p at once, one samples nodes over multiple, $T = \Omega(\log n)$, rounds, where in each round nodes are sampled with some smaller probability $p' \approx p/T$. This allows to study and analyze the emergence and merging of connected components as time progresses and more and more nodes are sampled.

Next, let us take a look at a single cut, the canonical bad cut consisting of a k-edge-matching as discussed in Observation 2. We emphasize that understanding the behavior of *all* cuts simultaneously is the part that makes the problem challenging, but focusing on this single cut should be sufficient for delivering the right intuition about the key new element in our analysis.

In the cut consisting of a k-edge-matching, in any round, both endpoints of an edge will become sampled with probability p'^2. Since there are k such edges, the probability that at least one edge gets sampled in a round is bounded by kp'^2. Now, in order for a cut to merge with high probability in this way over the course of T rounds, we need that $Tkp'^2 = kp^2/T = \Omega(\log n)$. Assuming $T = \Omega(\log n)$, this results in $p > \log n/\sqrt{k}$ being a necessary condition. This explains in a very simplified manner why the argument in [6] does not work for $p = o(\log n/\sqrt{k})$.

In this work, we refine this layer-by-layer sampling by further exploiting that connectivity evolves gradually. In particular, while the probability of obtaining one complete edge in one round is only p'^2, and thus quite small, the number of sampled nodes on each side of the cut grows by roughly kp' in each round. Thus, after λ/kp' rounds for some $\lambda = \Omega(\log n)$, the number of such neighbors is at least λ with high probability. Each of these nodes intuitively already goes half way in crossing the cut. In particular, with

λ such nodes, there is a chance of $\lambda p'$ per each of the next rounds to complete such a semi-sampled edge into a fully sampled edge that crosses the cut. This means after such λ-"semi-connectivity" is achieved, with high probability, only $\log n/\lambda p'$ further rounds are needed to get an edge crossing the cut. The optimal value for λ is now chosen to balance between the λ/kp' rounds to achieve it and the $\log n/\lambda p'$ additional rounds required to lead to connectivity. This leads to $\lambda = \sqrt{\log n/k}$, and results in $T = \sqrt{\log n/k}/p'$ rounds and a sampling probability of $p = \sqrt{\log n/k}$ being sufficient for a single cut.

In the above description we focused on a single cut. Understanding, however, the behavior of all (the exponentially many) cuts together turns out significantly more complex. Overall, the main technical challenge in this paper is to develop notions, definitions and arguments to prove that semi-connectivity indeed gets established quickly, for all cuts.

2.3 Proof of Theorem 6 via Semi-Connectivity. In this section, we present the formal definition of semi-connectivity, and explain how the proof of Theorem 6 incorporates the analysis of semi-connectivity. At a high level, the process of sampling consists of three parts (of unequal probability mass) for obtaining (i) domination, (ii) λ-semi-connectivity for a $\lambda = \Theta(\sqrt{k \log(n/\delta)})$, and (iii) connectivity. Establishing domination is trivial, and the proof of connectivity after having λ-semi-connectivity follows easily from the layer-by-layer analysis of [6]. The key challenge in our analysis is to prove λ-semi-connectivity given domination. Particularly, we show that $\Theta(\lambda/k)$ mass of sampling probability suffices to increases the semi-connectivity of a dominating set by an additive term of λ, for $\lambda = \Omega(\log n)$.

Basic Notation: We say that a node $u \in V$ is a *neighbor of* $S \subseteq V$ or *adjacent to* S if u is adjacent to some node $v \in S$ and $u \notin S$. The set of neighbors of S is denoted by ∂S. An edge/path *between two sets* S *and* S' is one with endpoints $u \in S$ and $u' \in S'$. For a set $S \subseteq V$, we use $G[S]$ to denote the subgraph of G induced by S.

Following the approach highlighted above, we formally define the notion of semi-connectivity as follows.

DEFINITION 7. (λ-SEMI-CONNECTED SET) *A node set $S \subseteq V$ is* λ-semi-connected*, for some $\lambda \geq 0$, if for any partition of S into two sets T and $S \setminus T$ with no edges between them, T and $S \setminus T$ have at least λ common neighbors, i.e., $|\partial T \cap \partial(S \setminus T)| \geq \lambda$.*

It is straightforward to see that if a set S is $(\lambda+1)$-semi-connected, then it is also λ-semi-connected. And any connected set is λ-semi-connected for any $\lambda \geq 0$, as the condition in Definition 7 is vacuously true in this case.

In the following claim, we observe that adding a node from V to a λ-semi-connected set $S \subseteq V$ does not break semi-connectivity, if S is a *dominating* set.

CLAIM 8. *If $S \subseteq V$ is a λ-semi-connected dominating set, then for any node $u \in V \setminus S$, the set $S \cup \{u\}$ is also λ-semi-connected.*

Proof. Suppose, for the sake of contradiction, that there is a partition of the set $S' = S \cup \{u\}$ into two sets T' and $S' \setminus T'$, such that these sets have no edges between them, and have fewer than λ common neighbors. We assume w.l.o.g. that $u \in T'$. We observe that $T' \neq \{u\}$, because S is a dominating set and thus if $T' = \{u\}$ then there would be an edge between T' and $S' \setminus T'$. Thus, the set $T = T' \setminus \{u\}$ is non-empty, and the two sets T and $S \setminus T = S' \setminus T'$ constitute a partition of S. We have that T and $S \setminus T$ have no edges between them, for otherwise the same edge would also connect T' and $S' \setminus T'$. However, T and $S \setminus T$ have fewer than λ common neighbors, as each common neighbor of T and $S \setminus T$ is also a common neighbor of T' and $S' \setminus T'$. This implies that S is *not* λ-semi-connected—a contradiction. ∎

We now show that it suffices to sample nodes with probability $\Theta(\log(n/\delta)/\lambda)$ to end up with a connected subgraph with probability $1-\delta$, if we start from an initial set S of sampled nodes that is a λ-semi-connected dominating set.

LEMMA 9. *Let $S \subseteq V$ be a λ-semi-connected dominating set. Sampling each node $u \in V \setminus S$ with probability $\log_\gamma(n/\delta)/\lambda$, where $\gamma = \frac{2e}{e+1}$, yields a set S' such that the graph $G[S \cup S']$ is connected with probability at least $1-\delta$.*

Proof. We perform sampling in rounds, where in each round every node that has not been selected yet is sampled with probability $1/\lambda$. The total number of rounds is $r = \log_\gamma(n/\delta)$, thus the probability for any given node $u \in V \setminus S$ to get selected is one of those rounds is at most $r/\lambda = \log_\gamma(n/\delta)/\lambda$, as required. Let S_i, for $0 \leq i \leq r$, denote the set consisting of all nodes selected in the first i rounds and all $u \in S$ (so $S_0 = S$). Further, let X_i denote the number of connected components of graph $G[S_i]$, and let $Y_i = X_i - 1$. We will now bound $\mathbf{E}[Y_i]$.

Observe that adding a new node u to a dominating set D yields another dominating set D' in which all connected components of $G[D]$ that u is adjacent to (if there are more than one) "merge" into a single connected component. Further, if D is λ-semi-connected then so is D', by Claim 8.

Fix a set S_i and suppose that $G[S_i]$ is disconnected, i.e., $X_i > 1$. From the above observations it follows that S_i is a λ-semi-connected dominating set. Hence, each connected component C of $G[S_i]$ has at least λ common neighbors with other connected components, and if any of those common neighbors gets selected in round $i+1$ then C gets merged with some component. Then the probability of C to get merged in round $i+1$ is at least $1-(1-1/\lambda)^\lambda \geq 1-1/e$. Since the drop $X_i - X_{i+1}$ in the number of connected

components in round $i+1$ is at least half the total number of connected components that get merged, it follows that

$$\mathbf{E}[X_i - X_{i+1} \mid S_i] \geq \frac{1-1/e}{2} X_i = (1-1/\gamma)X_i.$$

This inequality assumes that $X_i > 1$. To lift this assumption we consider the random variables $Y_i = X_i - 1$ instead. We have

$$\begin{aligned}\mathbf{E}[Y_i - Y_{i+1} \mid S_i] &= \mathbf{E}[X_i - X_{i+1} \mid S_i] \\ &\geq (1-1/\gamma)X_i \geq (1-1/\gamma)Y_i.\end{aligned}$$

The above inequality $\mathbf{E}[Y_i - Y_{i+1} \mid S_i] \geq (1-1/\gamma)Y_i$ also holds (trivially) when $X_i = 1$, since then $Y_i = 0$. Taking now the unconditional expectation yields $\mathbf{E}[Y_i - Y_{i+1}] \geq (1-1/\gamma)\,\mathbf{E}[Y_i]$, which implies $\mathbf{E}[Y_{i+1}] \leq \mathbf{E}[Y_i]/\gamma$. Applying this inequality repeatedly gives

$$\mathbf{E}[Y_i] \leq \mathbf{E}[Y_0]/\gamma^i \leq n/\gamma^i,$$

since $Y_0 < n$. Setting $i = r$ yields $\mathbf{E}[Y_r] \leq n/\gamma^r = \delta$, as $r = \log_\gamma(n/\delta)$. By Markov's inequality then we obtain $\Pr(Y_r > 0) = \Pr(Y_r \geq 1) \leq \mathbf{E}[Y_r]/1 \leq \delta$. Therefore, the probability that there is only one connected component at the end of the last round is at least $1-\delta$. ∎

Lemma 9 assumes that we start from some initial set S of already sampled nodes, which is a λ-semi-connected dominating set. In the next simple lemma, we show that it suffices to sample nodes with probability just $\Theta(\log(n/\delta)/k)$ to obtain a dominating set with probability $1-\delta$. Recall that k is the vertex connectivity of the graph.

LEMMA 10. *Sampling each node with probability $\ln(n/\delta)/k$ yields a dominating set with probability at least $1-\delta$.*

Proof. From the k-vertex-connectivity of the graph, it follows that each node has degree at least k. Thus the probability for a given node that none of its neighbors gets selected is at most $\left(1-\frac{\ln(n/\delta)}{k}\right)^k \leq e^{-\frac{\ln(n/\delta)}{k}\cdot k} = \delta/n$. By the union bound, the probability that this happens for at least one of the n nodes is at most δ. ∎

It remains to bound the sampling probability needed to achieve λ-semi-connectivity. This is the key part in our analysis. In particular, we show that a sampling probability of $\Theta((\lambda+\log n)/k)$ suffices to achieve λ-semi-connectivity. Section 2.4 is dedicated to the proof of this result, which is formally stated as follows.

LEMMA 11. (KEY SEMI-CONNECTIVITY CLAIM) *Let set $S \subseteq V$ be a dominating set. Sampling each node $u \in V \setminus S$ with probability $16\lambda/k$ yields a set S' such that $S \cup S'$ is a λ-semi-connected set with probability at least $1 - n/2^\lambda$.*

We now have all the ingredients to prove Theorem 6.

Proof of Theorem 6. If $k = O(\log(n/\delta))$ then the theorem holds trivially by choosing the constant β such that $\beta\sqrt{\log(n/\delta)/k} \geq 1$. Below we assume that $k > \log(3n/\delta)$.

We consider three phases. First we sample nodes with probability $\ln(3n/\delta)/k$, and obtain from Lemma 10 that the resulting set is dominating with probability $1-\delta/3$.

In the next phase, we sample the nodes not yet selected with probability $16\lambda/k$, for $\lambda = \sqrt{k\log(3n/\delta)}$. From Lemma 11 it follows that if the nodes selected in the first phase form a dominating set, then the nodes selected in the first two phases form a λ-semi-connected dominating set with probability $1 - n/2^\lambda$. We have $1 - n/2^\lambda \geq 1 - \delta/3$, because $\lambda = \sqrt{k\log(3n/\delta)} \geq \log(3n/\delta)$, as we have assumed that $k > \log(3n/\delta)$.

In the last phase, we sample the remaining nodes with probability $\log_\gamma(n/\delta)/\lambda$, and obtain from Lemma 9 that the probability for the subgraph induced by the nodes selected in the three phases to be connected is at least $1-\delta/3$, provided that the nodes selected in the first two phases are a λ-semi-connected dominating set.

A union bound over all three phases shows that the probability of ending up with a connected graph is indeed $1-\delta$, and the total sampling probability is at most

$$\frac{\ln(3n/\delta)}{k} + \frac{16\sqrt{k\log(3n/\delta)}}{k} + \frac{\log_\gamma(n/\delta)}{\sqrt{k\log(3n/\delta)}},$$

which is $O\big(\sqrt{\log(n/\delta)/k}\big)$. ∎

2.4 Proof of Lemma 11: Sampling Threshold for λ-Semi-Connectivity. Given a dominating set $S \subseteq V$, we prove that sampling each node $u \in V \setminus S$ with probability $16\lambda/k$ yields a set S' such that $S \cup S'$ is a λ-semi-connected set with probability at least $1 - n/2^\lambda$.

For the analysis we assume that sampling is carried out in *rounds*. In each round, each node that has not been selected yet is sampled with probability $1/k$. Within a round, the sampling of nodes is done sequentially, in *steps*, with a single node considered at each step (the order in which nodes are considered in a round can be arbitrary).

In the following we denote by S_t the set containing all nodes selected in the first t steps and all $u \in S$ (so $S_0 = S$).

We assume that at any point in time, each *edge* has a color from the set {black, gray, white, color-1, ..., color-λ}. We have the following coloring initially: Edges with both endpoints in S are black; the edges between each node $u \in V \setminus S$ and u's neighbors from S are gray; and all remaining edges (between nodes from $V \setminus S$) are white. There are no color-i edges initially, for any $1 \leq i \leq \lambda$. As sampling proceeds, edges may change color. The possible changes are that white edges may switch to color-i, for some i, and edges of any color may switch to black. At any point

in time we have the following invariants: An edge is black iff both its endpoints belong to S; if an edge is gray or of color-i, for some i, then exactly one of its endpoints is in S and the other in $V \setminus S$; if both endpoints of an edge are in $V \setminus S$ then this edge is white. (But it is possible for a white edge to have one endpoint in S and the other in $V \setminus S$.)

Intuitively, we will show that in the end, for each color i, all the connected components of the subgraph induced by the sampled nodes are connected by length-2 paths consisting of color-i edges and gray edges. Moreover, these paths can be chosen in such a way that the connector paths for different colors are internally vertex-disjoint so that together, they imply that the sampled set is λ-semi-connected.

Before we describe precisely the color changes that take place in a step we need to introduce some terminology.

DEFINITION 12. (i-NOVO-CONNECTIVITY) *A simple path between two* sampled *nodes is an* i-novo-path*, for some* $1 \leq i \leq \lambda$*, if (1) each edge in the path has a color from the set* $\{$*black, gray, color-*$i\}$*, and (2) for any two consecutive edges whose common endpoint is not sampled, at least one of them is a color-*i *edge. Two sampled vertices are* i-novo-connected *if there is an* i*-novo-path between them. Finally, an* i-novo-connected component*, or simply* i-novo-component*, is a maximal subset of the sampled nodes such that any two nodes in that set are* i*-novo-connected.*[6]

We describe now the color changes that take place at some step $t \geq 1$. Suppose that node $u \notin S_{t-1}$ is considered for sampling in step t. If u is not selected in that step, i.e., $S_t = S_{t-1}$, then the edge colors do not change. If u is selected, i.e., $S_t = S_{t-1} \cup \{u\}$, all edges uv with $v \in S_{t-1}$ become black, and then the following λ *sub-steps* are performed. In each sub-step $i = 1, \ldots, \lambda$, some edges incident to u may switch from white to color-i. Precisely, an edge uv switches to color-i if all the conditions below hold:

1. uv is white before sub-step i;
2. v is adjacent to only one i-novo-component before step t—we say v is an *exclusive* neighbor of that component;
3. u is not adjacent to the same i-novo-component as v before step t;

and we also have the additional rule that

4. if there are more than one node v that satisfy the three conditions above and are all adjacent to the *same* i-novo-component before step t, then only one edge uv is colored with color-i (choosing an arbitrary one among those nodes v).

[6]We will see, in Claim 13, that i-novo-connectivity is an equivalence relation between sampled nodes.

Next, we analyze how the i-novo-components evolve over time as the sampling proceeds. We first show that i-novo-connectivity is indeed an equivalence relation between sampled nodes, making the notion of an i-novo-component well defined.

CLAIM 13. *i-novo-connectivity is an equivalence relation between sampled nodes.*

Proof. It is straightforward to see that i-novo-connectivity is reflexive and symmetric. It remains to show transitivity, i.e., if a node u is i-novo-connected with nodes v and w, then v and w are i-novo-connected with each other.

Suppose, for the sake of contradiction, that the transitivity property is violated at some point, and let t be the earliest step when this happens. That is, at some point during step t, there is some i and nodes u, v, w such that u is i-novo-connected with both v and w, but v and w are not i-novo-connected with each other. Recall that before the first step there are no color-i edges, so at that time two nodes are i-novo-connected iff they are connected (by a path of black edges), and thus transitivity holds.

Let p be an i-novo-path between v and u, and q an i-novo-path between w and u. Let x be the first node where the two paths intersect when going from w towards u on path q. We define r to be the concatenation of the subpath of p connecting v and x and of the subpath of q connecting x and w. Note that r is a simple path connecting v and w, and node x is the only node of r that is in the intersection of paths p and q. Further note that x cannot be a sampled node because in that case the path r is an i-novo-path connecting nodes v and w and thus v and w are i-novo-connected. Hence, in particular, $x \notin \{u, v, w\}$. Let v' and w' be the neighbors of x in path r towards v and w, respectively, and let u' be the neighbor of x in p towards u. Note that it is possible that $u = u'$, $v = v'$, or $w = w'$. We also observe that both edges xv' and xw' must be gray, because if at least one of them is color-i then r is an i-novo-path.

We have thus established that node x is not selected and both edges xv' and xw' are gray. Since xv' and xu' are consecutive edges in i-novo-path p and x is not selected (i.e., sampled), it follows that xu' must be color-i. Consider the step $t' \leq t$ at which this edge changed from white to color-i, when u' was selected. It must be the case that before step t', and thus before step t, x was an exclusive neighbor to a single i-novo-component. We stress here that at any point before step t, i-novo-components are well defined as the transitivity property holds for i-novo-connectivity up to that step, because of the minimality of t. Since xv' and xw' are both gray and since no edge becomes gray at any step, these edges were also gray before step t', which means that v' and w' were in the same i-novo-component before step t.

We now argue that at least one of nodes v and w is also in the same i-novo-component as v' and w' before step t: The

subpath of r between v and v' and the subpath between w and w' are both i-novo-paths and they do not intersect. We also have that in step t, as in any step, only edges incident to the node selected in that step (if one is indeed selected) may change color. Since the subpaths above do not share a common node, at least one of them does not change in step t. Suppose, w.l.o.g., that the subpath between w and w' does not change. It follows then that w is in the same i-novo-component as v' and w' before step t. Thus there is an i-novo-path q_1 between w and v' before step t. We further denote the subpath of p between v and v' by p_1.

We now apply a very similar argument as above using paths p_1 and q_1 in place of p and q. Two key observations here are that (1) p_1 is a proper subpath of p, and thus p_1 is strictly shorter than p; and (2) the i-novo-path q_1 between w and v' existed before step t. Similarly to before, we let x_1 be the first node in the intersection of p_1 and q_1 when going from node v towards node v' on path p_1 and we denote r_1 the concatenation of the subpaths of p_1 and q_1 connecting x_1 with v and w, respectively. Again x_1 cannot be sampled as otherwise r_1 is an i-novo-path. Defining v'_1 and w'_1 in a similar manner as before and using an analogous argument, we get that v'_1 and w'_1 are in the same i-novo-component before step t. Observe that w is also in that i-novo-component, because w and w'_1 are connected by an i-novo-path before step t, namely the subpath of q_1 between w and w'_1 (see observation (2) made earlier). Thus there is an i-novo-path q_2 between w and v'_1 before step t. We then define p_2 to be the subpath of p_1 between v and v'_1, and repeat the exact same argument for p_2 and q_2, and so on. Since the length of the paths $p_1, p_2, \ldots$ strictly decreases, it follows that for some s we will have $v'_s = v$, and from this we obtain that v and w are in the same i-novo-component before step t—a contradiction. ∎

Two distinct i-novo-components *merge* into a single component, if an i-novo-path is created between them. In particular, the two components merge if a common neighbor of them is selected and thus a black path is created between them, or if a neighbor u of the one component is selected and then an edge uv to a neighbor v of the other component switches to color-i.

CLAIM 14. *Suppose that node u is selected in step t, and edge uv is colored with color-i in that step. If u belongs to i-novo-component C before sub-step i of step t, and v is adjacent to i-novo-component C' before step t, then coloring edge uv results in merging C and C'.*

Proof. We need to show that an i-novo-path is created between C and C'. Let w be a node in $C' \cap S$ to which v is connected (recall that S is the set of sampled nodes initially, before the first step). The node w exists since S is a dominating set so there must be a node in S which is connected to v, and that node has to be in C' because v is an exclusive neighbor of C'. Therefore, the edge wv must be colored gray since $w \in S$ and $v \in V \setminus S_t$. Since the edge vu is colored with color-i, this implies an i-novo-path between w and u, completing the proof. ∎

The next claim says that at any point, the partition of sampled nodes into i-novo-components is a refinement of the partition into $(i-1)$-novo-components.

CLAIM 15. *At any point in time, for any $2 \leq i \leq \lambda$, each i-novo-component is a subset of some $(i-1)$-novo-component.*

Proof. The proof is by induction on the number of steps t. The base case holds since when $t = 0$ there are only white, black, and gray edges, implying that any i-novo-component is also a j-novo-component for every $1 \leq i, j \leq \lambda$. Assume the claim holds after the first $t-1$ steps and consider step t.

Suppose node u is selected at step t, and let $v_1, \ldots, v_\ell \in S_{t-1}$ be the neighbors of u that are already selected. Since $S_{t-1} \supseteq S$ is a dominating set, we have that $\ell \geq 1$. Let $C_1, \ldots, C_\ell$ denote the i-novo-components to which $v_1, \ldots, v_\ell$, respectively, belong to before step t (these components are not necessarily distinct). When u is selected, all edges uv_j, for $1 \leq j \leq \ell$, become black, and a new i-novo-component $C = \{u\} \cup C_1 \cup \cdots \cup C_\ell$ is formed, replacing $C_1, \ldots, C_\ell$. Similarly, if $C'_1, \ldots, C'_\ell$ are the $(i-1)$-novo-components of $v_1, \ldots, v_\ell$, respectively, before step t, then a new $(i-1)$-novo-component $C' = \{u\} \cup C'_1 \cup \cdots \cup C'_\ell$ replaces $C'_1, \ldots, C'_\ell$. From the induction hypothesis it follows that $C_j \subseteq C'_j$ for any $1 \leq j \leq \ell$, and thus $C \subseteq C'$. Also, any other i-novo-component which was a subset of one of the C'_j before step t, is now a subset of C. This proves that the claim holds before the first sub-step of step t.

It is also immediate that the claim holds before sub-step i of step t, because in the first $i-1$ sub-steps i-novo-components do not change, and in step $i-1$, $(i-1)$-novo-components may merge, but merging existing $(i-1)$-novo-components cannot invalidate the claim.

Consider now sub-step i. Let uv be an edge that is white and turns color-i at this sub-step, and let w be a neighbor of v in the i-novo-component for which v is an exclusive neighbor before step t. We need to show that u and w are $(i-1)$-novo-connected at this time. We assume otherwise, towards a contradiction, and show that uv fulfilled all the requirements for becoming color-$(i-1)$ at sub-step $i-1$. First, uv was white at sub-step $i-1$ since it is white before sub-step i. Second, u and v were not adjacent to the same $(i-1)$-novo-component before step t, otherwise u and w belong to the same $(i-1)$-novo-component, which we assumed to be false. Finally, v was adjacent to only one $(i-1)$-novo-component before step t, since otherwise, by the induction hypothesis, it is also adjacent to more than one i-novo-component. Since uv remained white after sub-step

$i-1$, it must be that u and w were already $(i-1)$-novo-connected, or became $(i-1)$-novo-connected at this sub-step by a different common neighbor v', for which uv' became color-$(i-1)$.

Finally, in the remaining sub-steps of step t after sub-step i, i-novo-components and $(i-1)$-novo-components do not change and thus the claim holds. ∎

Next we define the notion of critical nodes of an i-novo-component, and show that each i-novo-component has at least k such critical nodes, where k is the vertex-connectivity of the graph. Further we show that the total number of j-novo-components for any $j \leq i$ that get merged in a round is bounded from below by the number of i-novo components for which some critical node is selected in that round.

DEFINITION 16. (CRITICAL NODES) *Let C be an i-novo-component before round r. A node u is* critical *for C in round r, if one of the following two conditions holds before round r: (1) u is a non-exclusive neighbor of C, i.e., it is adjacent to C and also to some i-novo-component $C' \neq C$; or (2) u is not in C or adjacent to C, but is adjacent to some exclusive neighbor of C.*

CLAIM 17. *Let C be an i-novo-component before round r. If there are more than one i-novo-components, then there are at least k critical nodes for C in round r.*

Proof. Since the graph is k-vertex-connected, there are k internally-disjoint paths connecting C to other i-novo-components. Moreover, since S is a dominating set, there are k such paths of length 2 or 3. This is because, from a longer path p, we can obtain a path p' of length 2 or 3 whose internal nodes are also internal nodes of p: If p contains a node x that is a non-exclusive neighbor of C then we let x be the internal node of p', and p' has length 2. If no such node x exists, we let $y, z \notin C$ be a pair of consecutive nodes in p such that y is a neighbor of C and z is not, and we let these two nodes be the internal nodes of p'—p' has length 3 in this case. Since z is not adjacent to C it must be adjacent to another i-novo-component, as S is a dominating set.

Consider now the internal nodes of all paths of length 2 between C and other i-novo-components. These nodes are critical for C, by the first condition of Definition 16. If the number ℓ of such nodes is at least k then we are done. So, suppose that $\ell < k$, and consider the paths of length 3 that are internally-disjoint to all paths of length 2. Each internal node y on such a path p that is adjacent to C is an exclusive neighbor of C, otherwise it would have been on a path of length 2. Therefore, the other internal node z on p will be adjacent to y and to a component other than C, thus it is critical for C by the second condition of Definition 16. Since there are at least $k-\ell$ such disjoint paths of length 3 with internal nodes that are not on paths of length 2, we have at least $k-\ell$ additional critical nodes for C. ∎

CLAIM 18. *The total number of j-novo-components, for any $j \leq i$, that get merged in round r is at least equal to the number of i-novo-components before round r, for which some critical node gets selected in a step of round r.*

Proof. Let C be an i-novo-component before round r, for which a critical node gets selected in some step of round r. Let t be the earliest step when this happens, and let u be the critical node of C selected at that step. Suppose that C does not merge with any other i-novo-component in round r. Precisely, there is no i-novo-component $D \neq C$ before round r, such that an i-novo-path is created between C and D during round r. We prove that a merge occurs between two j-novo-components in step t, for some $j < i$, and moreover, we have a distinct such merge for each such C.

First we observe that, from the assumption that C does not merge with another i-novo-component in round r, u cannot be a non-exclusive neighbor of C before round r, as this would imply that u was also adjacent to some other i-novo-component $D \neq C$ before round r, and thus selecting u creates a black path between C and D. Hence, from the definition of a critical node, we know that before the start of round r, u is not adjacent to C, but it is adjacent to some exclusive neighbor v of C, and it is also adjacent to some i-novo-component $D \neq C$ (since S is a dominating set). Consider now the graph before step t. Let $C' \supseteq C$ and $D' \supseteq D$ denote the current i-novo-components at that time, containing respectively C and D. (Note that C' may be a proper superset of C even though C does not merge, e.g., if some exclusive neighbor of C was selected in a previous step of the round.) Then u is a neighbor of D', and we claim that it is not a neighbor of C', otherwise selecting u would create an i-novo-path between C' and D', and thus one between C and D. We also claim that v is an exclusive neighbor of C': First, v cannot have been selected in an earlier step, because then selecting u would create a black path between C' and D'. Second, v cannot be a non-exclusive neighbor of C', otherwise some neighbor u' of v was selected before step t in round r, and u' does not belong to C', implying that u' is not a neighbor of C, and thus u' must be a critical node for C for round r. But this contradicts the assumption that u is the first critical node of C to get selected in round r.

We have thus established that before step t, u is adjacent to D' and not adjacent to C', and v is an exclusive neighbor of C'. We now argue that uv cannot be white before sub-step i: If uv were white, then Conditions 1–3 required for uv to switch to color-i would be satisfied. From this and the last condition, 4, it would follow that either uv or some other edge uv' would become color-i, for some exclusive neighbor $v' \neq v$ of C'. From Claim 14 then, this would create an i-novo-path between C' and D', and thus between C and D.

Since uv is not white before sub-step i (of step t), and it is clearly white at the beginning of step t as none of its endpoints is selected at the time, it follows that uv switches

from white to color-j, for some $j < i$, in sub-step j of step t. Claim 14 then implies that two j-novo-components R and R', such that $u \in R$ and v is an exclusive neighbor of R', get merged by coloring uv.

Finally, suppose that u is also a critical node for some i-novo-component $C'' \neq C'$, and that, like C', C'' does not merge with D'. As before this implies that a pair of j'-novo-components Q and Q' get merged instead, for some $j' < i$. We observe that $(j, R, R') \neq (j', Q, Q')$, i.e, the merges of C' with D' and of C'' with D' are prevented by distinct merges of j,j'-novo-components for smaller j and j'. This is clear if $j' \neq j$, and if $j' = j$ it follows from Condition 4 for coloring an edge color-j.

We have thus shown that a distinct merge of two j-novo-components occurs, for some $j \leq i$, for each i-novo-component C before round r for which a critical node gets selected in round r. ∎

We will use Claim 18 to show in the next claim that the expected drop in a round of the total number of j-novo-components for all $j \leq i$ is at least a linear function in the expected number of i-novo-components before the round.

For $1 \leq i \leq \lambda$ and $r \geq 0$, let $Y_{i,r}$ denote the number of i-novo-components after the first r rounds. Also let $X_{i,r} = Y_{i,r} - 1$, and let $x_{i,r} = \mathbf{E}[X_{i,r}]$.

CLAIM 19. *For any $1 \leq i \leq \lambda$ and $r \geq 1$, and for $\rho = \frac{e-1}{2e}$,*

$$\sum_{j=1}^{i}(x_{j,r-1} - x_{j,r}) \geq \rho x_{i,r-1}.$$

Proof. By Claim 17 we have that every i-novo-component C has at least k critical nodes. The probability that a node gets sampled in a given round is $1/k$, thus the probability at least one of C's critical nodes gets sampled in round r is at least $1-(1-1/k)^k \geq 1-1/e = 2\rho$. If this happens we say that C *causes a merge*.

By Claim 18 we have that the total number of j-novo-components over all $j \leq i$ that get merged in round r is at least equal to the number of i-novo-components that cause a merge. The decrease in the number of j-novo-components is at least half the number of those merges.

Let X_C be an indicator random variable that is 1 if component C causes a merge and 0 otherwise. Given the number $Y_{i,r-1}$ of i-novo-components before round r, the expected number of such components that cause a merge is $\mathbf{E}[\sum X_C | Y_{i,r-1}] = \sum \mathbf{E}[X_C | Y_{i,r-1}] \geq 2\rho X_{i,r-1}$. Then

$$\mathbf{E}\left[\sum_{j=1}^{i}(Y_{j,r-1} - Y_{j,r}) \,\middle|\, Y_{i,r-1}\right] \geq \frac{1}{2}\,\mathbf{E}\left[\sum X_C | Y_{i,r-1}\right] \geq \rho X_{i,r-1},$$

and by the law of total expectation

$$\mathbf{E}\left[\sum_{j=1}^{i}(Y_{j,r-1} - Y_{j,r})\right] \geq \rho\,\mathbf{E}[X_{i,r-1}].$$

We therefore obtain

$$\begin{aligned}\sum_{j=1}^{i}(x_{j,r-1} - x_{j,r}) &= \mathbf{E}\left[\sum_{j=1}^{i}(X_{j,r-1} - X_{j,r})\right] \\ &= \mathbf{E}\left[\sum_{j=1}^{i}(Y_{j,r-1} - Y_{j,r})\right] \\ &\geq \rho\,\mathbf{E}[X_{i,r-1}] \;=\; \rho x_{i,r-1},\end{aligned}$$

which concludes the proof. ∎

Using Claim 19 we establish the following upper bound on $x_{i,r}$.

CLAIM 20. *For any $1 \leq i \leq \lambda$ and $r \geq 0$, and for $\rho = \frac{e-1}{2e}$ as in Claim 19, we have*

$$x_{i,r} \leq n\left(1-\frac{\rho}{2}\right)^r \left(1+\frac{2}{\rho}\right)^{i-1}.$$

Proof. We prove the statement by induction on r. For $r = 0$, we have $x_{i,r} \leq n$ and thus the claimed inequality clearly holds for all $i \in \{1, \ldots, \lambda\}$.

For the induction step, we assume that the inequality holds for $x_{i,r-1}$ for all $i \in \{1, \ldots, \lambda\}$, for some $r \geq 1$, and bound $x_{i,r}$. Solving the inequality in Claim 19 for $x_{i,r}$ and using the trivial lower bound $x_{j,r} \geq 0$ for all $j \leq i-1$, gives

$$x_{i,r} \leq (1-\rho)x_{i,r-1} + \sum_{j=1}^{i-1} x_{j,r-1}.$$

Applying the induction hypothesis to all terms on the right-hand side, we obtain

$$\begin{aligned} x_{i,r} &\overset{\text{(I.H.)}}{\leq} n\left(1-\frac{\rho}{2}\right)^{r-1} \\ &\quad \cdot \left[(1-\rho)\left(1+\frac{2}{\rho}\right)^{i-1} + \sum_{j=1}^{i-1}\left(1+\frac{2}{\rho}\right)^{j-1}\right] \\ &< n\left(1-\frac{\rho}{2}\right)^{r-1} \\ &\quad \cdot \left[\left(1+\frac{2}{\rho}\right)^{i-1}\left(-\rho + \sum_{h=0}^{\infty}\left(1+\frac{2}{\rho}\right)^{-h}\right)\right] \\ &= n\left(1-\frac{\rho}{2}\right)^{r-1}\left[\left(1+\frac{2}{\rho}\right)^{i-1}\left(1-\frac{\rho}{2}\right)\right], \end{aligned}$$

and thus the claim follows. ∎

Using Claim 20 and Markov's inequality we bound the number of rounds before there is just a single λ-novo-component left.

CLAIM 21. *All λ-novo-components have merged into a single component after 16λ rounds, with probability at least $1 - n/2^{\lambda}$.*

Proof. The probability that there is more than one λ-novo-component left after the first r rounds is $\Pr(Y_{\lambda,r} > 1) = \Pr(X_{\lambda,r} > 0) = \Pr(X_{\lambda,r} \geq 1) \leq \mathbf{E}[X_{\lambda,r}]/1$, by Markov's inequality. Also from Claim 20,

$$\mathbf{E}[X_{\lambda,r}] \leq n \left(1 - \frac{\rho}{2}\right)^r \left(1 + \frac{2}{\rho}\right)^{\lambda}.$$

Thus, in order to have $\Pr(Y_{\lambda,r} > 1) \leq n/2^{\lambda}$, it suffices that

$$n \left(1 - \frac{\rho}{2}\right)^r \left(1 + \frac{2}{\rho}\right)^{\lambda} \leq n/2^{\lambda}.$$

Solving for r and substituting ρ's value, $\rho = \frac{e-1}{2e}$, we obtain $r \geq \lambda \ln\left(\frac{\rho}{2\rho+4}\right) / \ln\left(1 - \frac{\rho}{2}\right) \approx 15.6085 \cdot \lambda$. ∎

We now show that if there is just one i-novo-component after step t, then the set S_t of nodes that have been selected by that time is i-semi-connected, i.e., for any partition of S_t into two sets T and $S_t \setminus T$ with no edges between them, the two sets have at least i common neighbors.

CLAIM 22. *If there is only one i-novo-component after step t, then for any partition of S_t into two sets T and $S_t \setminus T$ with no edges between them, there is a j-novo-path between T and $S_t \setminus T$ for each $1 \leq j \leq i$, such that all these paths have length 2 and are internally disjoint.*

Proof. We show by induction on j that for each $1 \leq j \leq i$ we can identify a j-novo-path of length 2 between T and $S_t \setminus T$ which is internally disjoint from all previous paths.

We first note that for every j, at any point during the random process, and for any subset T' of the set S' of nodes selected by that time, if T' and $S' \setminus T'$ are not connected by an edge and if there exists a j-novo-path between T' and $S' \setminus T'$, then there also exists such a j-novo-path of length 2. That is, there exists a path uwv, where $u \in T'$, $v \in S' \setminus T'$, and $w \notin S'$, and where at least one of the edges uw and vw is color-j and the other edge is either color-j or gray. This follows directly from the definition of j-novo-paths. As j-novo-paths only consist of color-j edges and of gray edges, at least one of the nodes of each edge has to be sampled and therefore on a j-novo-path, at least every second node has to be sampled. Any minimal j-novo-path connecting T' and $S' \setminus T'$ therefore has to be of length 2.

Since there is just a single i-novo-component after step t, Claim 15 gives that there is also just a single j-novo-component. Hence, there must be at least one j-novo-path connecting T and $S_t \setminus T$. In particular, using the above observation, there must be such a j-novo-path of length 2.

We need to show that among these length 2 j-novo-paths there is at least one that is internally disjoint from all of the ℓ-novo-paths given by the induction hypothesis, for all $1 \leq \ell < j$. Consider the first time in which such a length 2 j-novo-path is created. At this time one of the two edges of the path becomes color-j (the other edge is gray or became color-j in an earlier step). Let $u \in T$ and $v \in S_t \setminus T$ be the two endpoints of the path and $w \in V \setminus S_t$ be its internal node. Assume w.l.o.g., edge wv is the one that becomes color-j.

First we argue that there is no gray edge wv' between w and some node $v' \in S_t \setminus T$: Suppose there is such a gray edge wv'. Then edge uw cannot be color-j, because then uwv' is a j-novo-path created before uwv. Thus edge uw is gray. However, it must be the case that before wv became color-j, node w was an exclusive neighbor of a j-novo-component, and since both uw and wv' are gray, it follows that u and v' belonged to the same j-novo-component. Thus before wv became color-j there was already a j-novo-path between nodes $u \in T$ and $v' \in S_t \setminus T$, and thus there was also a j-novo-path of length 2 (see above) between two nodes from these two sets. This contradicts that uwv was the first such j-novo-path.

We now show that path uwv is internally disjoint from the ℓ-novo-paths given by the induction hypothesis for all $1 \leq \ell < j$. Fix some $\ell \in \{1, \ldots, j-1\}$. Assume, that there is an ℓ-novo-path of length 2 between T and $S_t \setminus T$ whose internal node is w; let $u''wv''$ denote that path where $u'' \in T$ and $v'' \in S_t \setminus T$. Since we have shown that there is no gray edge between w and some node from $S_t \setminus T$, it must be that edge wv'' is color-ℓ. We argue that v'' is selected *after* v: Suppose, for contradiction, that v'' is selected before v. Then when v is selected, v'' and u must be in the same j-novo-component, otherwise w is adjacent to two distinct j-novo-components, preventing wv from becoming color-j. Thus before wv became color-j there was a j-novo-path between nodes $u \in T$ and $v'' \in S_t \setminus T$, and thus there was also a j-novo-path of length 2 between two nodes from these two sets. This contradicts that uwv was the first such j-novo-path. We conclude that v'' was selected after v. This means that the ℓ-novo-path created when edge wv'' was colored, cannot be the first one created between T and $S_t \setminus T$, because that first path must have been created no later than the first j-novo-path, by Claim 15. ∎

By Claim 21, the sampling procedure results in a single λ-novo-component after at most 16λ rounds, with probability at least $1 - n/2^{\lambda}$. In each round the sampling probability is $1/k$, thus the total sampling probability is at most $16\lambda/k$. Once there is just a single λ-novo-component, by Claim 22 we have that the set S_t of sampled nodes is λ-semi-connected. This concludes the proof of Lemma 11.

3 Proof of Observations 2 and 3

In this section we prove the two statements from Section 1.2 that demonstrate the optimality of the bounds in Theorem 1.

Proof of Observation 2. The edge connectivity of G is at most k as it contains an edge-cut of size k, and thus its vertex connectivity is also at most k. On the other hand, it is easy to verify that the removal of any $k-1$ vertices does not disconnect G. Therefore G has vertex connectivity exactly k.

Let K denote the number of edges in the matching that survive after sampling (i.e., both their endpoint nodes are selected). The expected value of K is $\mathbf{E}[K] = kp^2$, since each edge survives with probability p^2. If $K \neq 0$ then K is an upper bound on the edge connectivity and thus on the vertex connectivity of the sampled subgraph. If $K = 0$ then it is still possible for the vertex connectivity to be positive, if no nodes are sampled from the one clique and at least one is sampled from the other. Let N_i, for $i = 1, 2$, denote the number of nodes selected in each of the two cliques respectively, and let Z_i be the indicator random variable with $X_i = 1$ if $N_i = 0$ and $X_i = 0$ otherwise. Then $\mathbf{E}[N_i] = pn$, and $\mathbf{E}[Z_i] = \Pr(N_i = 0) = (1-p)^n$. From the discussion above it follows that the vertex expansion of the sampled subgraph is at most $K + Z_2N_1 + Z_1N_2$, and thus the expected vertex expansion is at most

$$\begin{aligned}\mathbf{E}[K + Z_2N_1 + Z_1N_2] &= \mathbf{E}[K] + 2\,\mathbf{E}[Z_2N_1] \\ &= \mathbf{E}[K] + 2\,\mathbf{E}[Z_2] \cdot \mathbf{E}[N_1] \\ &= kp^2 + 2np(1-p)^n \\ &\leq kp^2 + 2npe^{-np}.\end{aligned}$$

If $p \geq 2\ln n/n$, then the second term in the last line above is $kp^2 \cdot (2n/kp)e^{-np} \leq kp^2 \cdot (1/k\ln n) = o(kp^2)$; thus the expected vertex connectivity is at most $kp^2 + o(kp^2)$.

For the probability that the sampled subgraph is disconnected, we first observe that if $p = O(1/n)$ then the subgraph is empty (and thus by convention disconnected) with constant probability. Thus, below we assume that $p \geq 2/n$. The probability that the sampled subgraph is disconnected is bounded from below by

$$\begin{aligned}&\Pr(K = 0 \wedge N_1 \neq 0 \wedge N_2 \neq 0) \\ &\quad\geq 1 - (\Pr(K \neq 0) + \Pr(N_1 = 0) + \Pr(N_2 = 0)) \\ &\quad= \Pr(K = 0) - 2\Pr(N_1 = 0) \\ &\quad= (1-p^2)^k - 2(1-p)^n.\end{aligned}$$

The second term in the last line is at most $(1-p^2)^k/2$, as

$$\frac{2(1-p)^n}{(1-p^2)^k} \leq \frac{2(1-p)^n}{(1-p^2)^n} = \frac{2}{(1+p)^n} \leq \frac{2}{(1+2/n)^n} \leq \frac{1}{2}.$$

It follows that the probability of the sampled subgraph to be disconnected is at least $(1-p^2)^k/2$, and this is at least $1/n^{o(1)}$ if $p = o(\sqrt{\log n/k})$. ∎

Proof Sketch of Observation 3. Since $p = \omega(1/n)$, we have with probability $1 - o(1)$ that at least one node gets selected from the first $n/3k$ cliques, and at least one gets selected from the last $n/3k$ cliques. The probability that no edge survives in the cut between two given consecutive cliques is $(1-p^2)^k = e^{-o(\log(n/k))} = \omega(k/n)$, as $p = o(\sqrt{\log(n/k)/k})$. Thus, the probability that at least one of the cuts between the middle $n/3k$ cliques gets disconnected is at least

$$1 - (1 - \omega(k/n))^{k/6k} = 1 - o(1),$$

where for this computation we just considered every second cut, i.e., $n/6k$ cuts in total, and used the fact that these cuts are vertex-disjoint. Combining the above yields the claim. ∎

4 Discussion

In this paper we show that when independently sampling vertices of a k-vertex-connected n-node graph with probability $p = \Omega(\sqrt{\log n/k})$, the remaining subgraph is connected and, moreover, has a vertex connectivity of $\Omega(kp^2)$, with high probability. Our proof is based on considering sampling as a gradual random process, and carefully analyzing the growth of (novo-)connected components using the new notions of semi-connectivity and novo-connectivity.

The constant factor hidden in the $\Omega(kp^2)$ asymptotic bound for the remaining vertex-connectivity obtained from our analysis is much smaller than 1. We leave open whether the remaining vertex-connectivity is in fact at least $kp^2(1-\epsilon)$, for an arbitrary small $\epsilon > 0$, assuming kp^2 is large enough, e.g., $kp^2 = \Omega(\log n/\operatorname{poly}(\epsilon))$. In particular, for a sampling probability of $p = 1 - o(1)$, or equivalently a sub-constant deletion probability, this would imply a remaining connectivity of $k - o(k)$ instead of just $O(k)$.

As stated in Corollary 4, our result implies that any graph can be partitioned into $\Omega(k/\log^2 n)$ vertex-disjoint connected dominating sets. While this is an improvement over the best previously known lower bound of $\Omega(k/\log^5 n)$, a logarithmic gap still remains compared to the upper bound of $O(k/\log n)$ for the number of vertex-disjoint connected dominating sets that is known for some graphs. Closing this gap is an intriguing open question for further research.

References

[1] N. Alon. A note on network reliability. In *Discrete Probability and Algorithms*, pages 11–14. Springer, 1995.

[2] A. A. Benczúr and D. R. Karger. Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time. In *Proc. 28th ACM Symposium on Theory of Computing (STOC)*, pages 47–55, 1996.

[3] B. Bollobás. *Random graphs*. Springer, 1998.

[4] B. Bollobás and O. Riordan. *Percolation*. Cambridge University Press, 2006.

[5] K. Censor-Hillel, M. Ghaffari, and F. Kuhn. Distributed connectivity decomposition. In *Proc. 32nd ACM Symposium on Principles of Distributed Computing (PODC)*, pages 156–165, 2014.

[6] K. Censor-Hillel, M. Ghaffari, and F. Kuhn. A new perspective on vertex connectivity. In *Proc. 25th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 546–561, 2014.

[7] A. Kanevsky. On the number of minimum size separating vertex sets in a graph and how to find all of them. In *Proc. 1st ACM-SIAM Symposium on Discrete Algorithm (SODA)*, pages 411–421, 1990.

[8] D. R. Karger. Global min-cuts in $\mathcal{RNC}$, and other ramifications of a simple mincut algorithm. In *Proc. 4th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 21–30, 1993.

[9] D. R. Karger. Random sampling in cut, flow, and network design problems. In *Proc. 26th ACM Symposium on Theory of Computing (STOC)*, pages 648–657, 1994.

[10] D. R. Karger. Using randomized sparsification to approximate minimum cuts. In *Proc. 5th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 424–432, 1994.

[11] D. R. Karger. A randomized fully polynomial time approximation scheme for the all terminal network reliability problem. In *Proc. 27th ACM Symposium on Theory of Computing (STOC)*, pages 11–17, 1995.

[12] D. R. Karger and M. S. Levine. Finding maximum flows in undirected graphs seems easier than bipartite matching. In *Proc. 30th ACM Symposium on Theory of Computing (STOC)*, pages 69–78, 1998.

[13] M. V. Lomonosov and V. P. Polesskii. Lower bound of network reliability. *Problems of Information Transmission*, 8:118–123, 1972.

[14] C. S. J. A. Nash-Williams. Edge-disjoint spanning trees of finite graphs. *Journal of the London Mathematical Society*, 36(1):445–450, 1961.

[15] D. A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proc. 36th ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2004.

[16] W. T. Tutte. On the problem of decomposing a graph into n connected factors. *Journal of the London Mathematical Society*, 36(1):221–230, 1961.

[17] A. Zehavi and A. Itai. Three tree-paths. *Journal of Graph Theory*, 13(2):175–188, 1989.